ICES REPORT 11-24

July 2011

Wavenumber Explicit Analysis for a DPG Method for the Multidimensional Helmholtz Equation

by

L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli

The Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Reference: L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli, Wavenumber Explicit Analysis for a DPG Method for the Multidimensional Helmholtz Equation, ICES REPORT 11-24, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, July 2011.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Wavenumber Explicit Analysis for a DPG Method for the Multidimensional Helmholtz Equation	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Texas at Austin,Institute for Computational Engineering and Sciences,Austin,TX,78712**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We study the properties of a novel discontinuous Petrov Galerkin (DPG) method for acoustic wave propagation. The method yields Hermitian positive de nite matrices and has good pre-asymptotic stability properties. Numerically, we nd that the method exhibits negligible phase errors (otherwise known as pollution errors) even in the lowest order case. Theoretically, we are able to prove error estimates that explicitly show the dependencies with respect to the wavenumber !, the mesh size h, and the polynomial degree p. But the current state of the theory does not fully explain the remarkably good numerical phase errors. Theoretically, comparisons are made with several other recent works that gave wave number explicit estimates. Numerically, comparisons are made with the standard nite element method and its recent modi cation for wave propagation with clever quadratures. The new DPG method is designed following the previously established principles of optimal test functions. In addition to the nonstandard test functions, in this work, we also use a nonstandard wave number dependent norm on both the test and trial space to obtain our error estimates.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**27**	

WAVENUMBER EXPLICIT ANALYSIS OF A DPG METHOD FOR THE MULTIDIMENSIONAL HELMHOLTZ EQUATION

L. DEMKOWICZ, J. GOPALAKRISHNAN, I. MUGA, AND J. ZITELLI

Abstract. We study the properties of a novel discontinuous Petrov Galerkin (DPG) method for acoustic wave propagation. The method yields Hermitian positive definite matrices and has good pre-asymptotic stability properties. Numerically, we find that the method exhibits negligible phase errors (otherwise known as pollution errors) even in the lowest order case. Theoretically, we are able to prove error estimates that explicitly show the dependencies with respect to the wavenumber ω, the mesh size h, and the polynomial degree p. But the current state of the theory does not fully explain the remarkably good numerical phase errors. Theoretically, comparisons are made with several other recent works that gave wave number explicit estimates. Numerically, comparisons are made with the standard finite element method and its recent modification for wave propagation with clever quadratures. The new DPG method is designed following the previously established principles of optimal test functions. In addition to the nonstandard test functions, in this work, we also use a nonstandard wave number dependent norm on both the test and trial space to obtain our error estimates.

Key words: time harmonic wave propagation; robustness; phase error; dispersion; high frequency; Petrov Galerkin

1. Introduction

The purpose of this paper is to introduce a Discontinuous Petrov Galerkin (DPG) method for the Helmholtz equation. We analyze the method and give error estimates with constants whose dependence on the wavenumber are explicitly shown. We also report results from many numerical experiments and numerically compare the performance of the DPG method with other methods. Although our theory predicts the same h convergence rates for the DPG method and the standard FEM, the numerical performance of the DPG method is far superior for high wave numbers. A striking numerical observation for which we do not have a theoretical explanation, is that the DPG method exhibits negligible phase errors.

The purpose of performing a wavenumber explicit analysis is to track down pollution errors and gain a better understanding of how they originate. These errors are well recognized as posing severe challenges in numerical simulation of wave propagation [2]. For many model problems, the pollution is manifested as phase errors which typically accumulate in the direction of wave propagation over the computational domain. Thus the concepts of pollution error, phase error, and discrete wave numbers are all closely related. To explain the pollution error in the context of finite element methods, we

Demkowicz and Gopalakrishnan gratefully acknowledge the collaboration opportunities provided by the IMA (Minneapolis) during their 2010-11 annual program. This work was supported in part by the AFOSR under FA9550-09-1-0608, the NSF under grant DMS-1014817, the FONDECYT project 1110272, and an ONR Graduate Traineeship.

follow [18]: Given that the exact solution u lies in a space U normed by $\|\cdot\|_U$, and the discrete solution u_h lies in an approximation subspace $U_h \subset U$, one observes that

$$\frac{\|u-u_h\|_U}{\|u\|_U} \leq C(\omega) \inf_{w_h \in U_h} \frac{\|u-w_h\|_U}{\|u\|_U}, \tag{1}$$

where ω is the wavenumber, $C(\omega) = C_1 + C_2\omega^\beta(\omega h)^\gamma$, and h is the element size. The infimum on the right measures the relative best approximation error, which is typically controlled when ωh is small, i.e., when enough elements per wavelength are used. However, the ω-dependence in $C(\omega)$, as measured by β, is a reflection of the pollution errors. For most standard methods of fixed order, the exponent β is observed to be a positive constant.

In contrast, we are able to show that for the new DPG method, quasi-optimality estimate (1) holds with $C(\omega)$ *independent of* ω. To be fully accurate, let us add that we will prove so for an "ideal" DPG method. The "practical" DPG method is a simple modification of the ideal DPG method (both methods will be introduced in Section 2). Note that the independence of $C(\omega)$ with respect to ω does not imply, in theory, that the phase errors vanish. This is because the U-norm may still contain ω-dependent terms. Yet, by performing a wavenumber explicit analysis, we take the first step towards understanding why we do not see phase errors in our numerical experiments. We fully analyze the ideal DPG method and provide error bounds explicit in the wavenumber ω, the mesh size h, and the finite element polynomial degree p.

By doing so, we are also able to compare our work with a few recent works which also give similar wavenumber explicit estimates. (These comparisons are in § 3.2.3.) For perspective, let us recall the famous negative result of [2] on the inevitability of pollution errors. Specifically, Babuška and Sauter [2] worked in the context of a standard method using a nine-point stencil (comparable to the standard FEM with bilinear elements). One then wonders what happens on more general meshes and methods. It has long been known that high order finite elements (with obviously larger stencils) do reduce pollution errors. However, a precise statement of this fact was only recently obtained by Melenk and Sauter [22]. We will compare our result with this work. There have been important recent developments in DG methods for the Helmholtz equation [12, 13, 16, 23]. Ultraweak formulations, ever since the works of [4, 17], have shown great potential in numerical solution of the Helmholtz equation, especially those approximations based on plane waves. Most modern DG methods are derived from ultraweak variational formulations. It is not surprising therefore that a fertile line of attack has been the use of plane waves within DG trial spaces. New theoretical tools improving the understanding of such methods have just been developed in the Plane Wave DG (PWDG) method [16]. We will compare our results to theirs. Another method we will compare with is an interior penalty method with complex stabilization developed in [12, 13]. They also provide an analysis of error terms characterizing explicitly the dependencies in wavenumber.

That our method is a Petrov-Galerkin method distinguishes it from all the above mentioned works. The guiding principle in designing Petrov-Galerkin methods is that one chooses trial spaces for good approximation properties, but one designs test spaces to obtain good stability properties. We have exemplified this principle in our earlier works [7, 8, 9, 25]. In [25], we considered the one-dimensional Helmholtz equation and obtained (1) with $C(\omega)$ independent of ω. Unfortunately however, we were not able to generalize the techniques there to the multi-dimensional case. In this paper, we will provide a different way of theoretical analysis that proves (1) for higher dimensions with

$C(\omega)$ independent of ω. The new analytical technique is built on the approach developed in [6].

We should note that there are *two* problems that usually cripple standard numerical methods for the Helmholtz equation: (i) the growth of the pollution error with frequency and (ii) the poor approximation of highly oscillatory wave solutions by polynomials. One can argue that better *trial* spaces are needed to overcome the latter. Recent works such as [16] raise the hope of better trial spaces. However, this is not the subject of this paper. We concentrate on overcoming the first difficulty by designing better *test* spaces. Nonetheless, it is important to note that any new developments in approximation spaces can be built into our approach easily. Indeed, our method computes test spaces that pair with any given trial approximation space.

Finally, let us remark on a practically attractive feature of our DPG method: It yields *Hermitian and positive definite* linear systems although the original Helmholtz operator is indefinite. This is not a surprise if one views the DPG method as a least squares method. Indeed, the DPG method is a least squares method in a nonstandard inner product, as clarified in our earlier papers (see e.g., [8, eq. (2.13)]). There have been other well known least squares methods, such as FOSLS, for solving the Helmholtz equation, notably [20]. They show a wavenumber-independent stability result if one stays in a subspace sufficiently away from resonant modes (although it is not clear how one may manage this numerically). They also claim reduced pollution, but the number of mesh points they use is far higher than what we use to obtain similar accuracy. More interesting are the results they give on multigrid solvers for the least square systems using the ideas of [3]. We hope to borrow these ideas to design efficient solvers for the DPG method, an issue for future research.

The paper is organized as follows. We first describe the DPG method, in the abstract, as well as for the Helmholtz application, in Section 2. In Section 3, we present our main result, namely the wave number independent error estimate of Theorem 3.1. A number of comparisons with other recent works are also made in this section. In Section 4, we prove Theorem 3.1. In Section 5 we present the results of numerical experiments illustrating the theoretical results and comparing the DPG method to other standard methods.

2. The ideal and the practical DPG methods

In this section we present the new DPG method. We begin by summarizing the DPG framework developed in [7, 8, 9, 25] in § 2.1. We then apply it to the Helmholtz setting. The method we are able to fully analyze is the method presented in § 2.2. The practically implemented method is described in § 2.3.

2.1. The abstract setting.

Let U (the "trial" space) and V (the "test" space) be vector spaces over the complex field $\mathbb{C}$, and let $b(\cdot,\cdot) : U \times V \mapsto \mathbb{C}$ be a continuous sesquilinear form. We assume that U is a reflexive Banach space under the norm $\|\cdot\|_U$ and that V is a Hilbert space under an inner product $(\cdot,\cdot)_V$ with a corresponding norm $\|\cdot\|_V$. The following assumption requires special mention, and we will verify it for the Helmholtz application later.

Assumption 1 (Injectivity). We assume that

$$\{w \in U : \ b(w,v) = 0, \ \forall v \in V\} = \{0\}. \tag{2}$$

Now, suppose we are given a continuous conjugate linear form $l(v)$ on V (we use standard terminology – see e.g., [24]). The variational problem we wish to approximate is:

$$\begin{cases} \text{Find } u \in U \text{ such that} \\ b(u,v) = l(v), \quad \forall v \in V. \end{cases} \tag{3}$$

To describe the DPG method for this abstract variational problem, we define

$$\|v\|_{\text{opt},V} = \sup_{w \in U} \frac{|b(w,v)|}{\|w\|_U} \tag{4}$$

and place one more assumption.

Assumption 2 (Norm equivalence). There are positive constants C_1, C_2 such that

$$C_1 \|v\|_V \leq \|v\|_{\text{opt},V} \leq C_2 \|v\|_V, \qquad \forall v \in V. \tag{5}$$

Clearly, this assumption implies that $\|v\|_{\text{opt},V}$ is a norm on V. This is called the *optimal test space norm* for reasons explained in [25]. Note that the norms $\|v\|_{\text{opt},V}$ and $\|v\|_V$ are not equal in general.

The approximate solution of the DPG method lies in a finite dimensional trial subspace $U_h \subset U$. The test space that pairs with the trial space U_h is a subspace $V_h \subset V$ that we now define: First, define the *trial-to-test operator* $T : U \mapsto V$ by

$$(Tu, v)_V = b(u,v), \qquad \forall v \in V. \tag{6}$$

Then the discrete test space is set by

$$V_h = T(U_h). \tag{7}$$

2.2. **The ideal DPG method.** The DPG approximation $u_h \in U_h$ satisfies

$$b(u_h, v_h) = l(v_h) \qquad \forall v_h \in V_h, \tag{8}$$

where V_h is as defined in (7). This is a Petrov-Galerkin type formulation as U_h and V_h are not generally identical. Next, we note two basic properties of this method.

The first property is that the stiffness matrix of the method is Hermitian and positive definite. Indeed, if $\{e_i\}$ is a basis for U_h, then setting $t_j = Te_i$, we find that the (i,j)-th entry of the stiffness matrix, namely B_{ij}, is

$$B_{ij} = b(e_i, t_j) = (t_i, t_j)_V = \overline{(t_j, t_i)}_V = \overline{b(e_j, t_i)} = \overline{B}_{ji}.$$

Thus the matrix is Hermitian. To see that it is also positive definite, let c be a complex vector and $w = c_1 e_1 + c_2 e_2 + \ldots$ be a basis expansion of any $w \in U_h$. Then, by (6),

$$c^* B c = b(w, Tw) = ||Tw||_V^2. \tag{9}$$

Now, by Assumption 1, it is obvious that T is injective. By (9), $c^* B c = 0$ if and only if $Tw = 0$, which hold if and only if c is the zero vector. As a consequence, we can use powerful iterative solvers for positive definite systems (like the conjugate gradient method) to obtain the DPG solution, even though the original Helmholtz operator is indefinite.

The second property is a basic convergence result for the abstract method. It is also easy to prove (see [25, Theorem 2.1]), but we omit the proof and simply state it here.

THEOREM 2.1: *Suppose Assumptions 1 and 2 hold. Then problems* (3) *and* (8) *are well-posed, and their respective solutions* u *and* u_h *satisfy the quasi-optimality estimate:*

$$\|u-u_h\|_U \leq \frac{C_2}{C_1}\inf_{w_h\in U_h}\|u-w_h\|_U.$$

2.3. **The practical DPG method.** In view of the fact that the test space V_h in (7) is determined by T, it is natural to ask if this is computationally feasible. As we shall see, the saving grace of the DPG formulation is that T is a local operator and consequently inexpensive to approximate. Yet, despite its locality, one must approximate the infinite dimensional T by a local finite dimensional operator $\tilde{T}$ in a computer implementation. I.e., in place of T, we use $\tilde{T}: U \mapsto \tilde{V}$ defined by

$$(\tilde{T}u,\tilde{v})_V = b(u,\tilde{v}), \qquad \forall\, \tilde{v}\in\tilde{V}, \tag{10}$$

where $\tilde{V}$ is the finite dimensional subspace of V. We will detail our specific choice of $\tilde{T}$ and $\tilde{V}$ for the Helmholtz application in Section 5. This perturbation of the ideal DPG method can be analyzed using the recently developed techniques in [15] and will not be discussed in this paper. Nevertheless, it is easy to verify that the stiffness matrix of this practical method is also Hermitian and positive-semidefinite.

2.4. **Application to the Helmholtz equation.** We consider a bounded domain $\Omega\subset\mathbb{R}^n$ $(n\geq 2)$ with Lipschitz boundary. Let $f\in L^2(\Omega)$ and $g\in H^{-1/2}(\partial\Omega)$. We consider the time-harmonic wave propagation problem as a first order system. A physically "right" way to do this is via the physics of *acoustical disturbances* [5]. Linearizing the isentropic Euler equations around a hydrostatic solution and assuming harmonic time variations, we obtain

$$\hat{\imath}\omega\vec{u} + \vec{\nabla}\phi = \vec{0}, \qquad \text{on } \Omega \tag{11a}$$

$$\hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{u} = f, \qquad \text{on } \Omega \tag{11b}$$

$$\vec{u}\cdot\vec{n} - \phi = g, \qquad \text{on } \partial\Omega, \tag{11c}$$

where $\vec{u}$ and ϕ are velocity and pressure variables, respectively, associated to the acoustic perturbations from equilibrium. Observe that taking the divergence of (11a) and substituting $\vec{u}$, we recover the usual second order form of the Helmholtz equation:

$$\begin{aligned} -\Delta\phi - \omega^2\phi &= \hat{\imath}\omega f && \text{on } \Omega \\ \frac{\partial\phi}{\partial n} + \hat{\imath}\omega\phi &= -\hat{\imath}\omega g && \text{on } \partial\Omega. \end{aligned}$$

Let Ω_h be a disjoint partitioning of Ω into open elements K such that $\overline{\Omega} = \cup_{K\in\Omega_h}\overline{K}$. We multiply the first two equations (11a) and (11b) by test functions and integrate by parts element-wise to obtain an ultraweak DG variational formulation. The details of the derivation are very similar to the case of the Poisson equation [6], so we omit them and simply present the DPG weak formulation, after a foreword on notations. Let $(\cdot,\cdot)_D$ denote the (sesquilinear) $L^2(D)$ inner product on any domain D. The notation $\langle\cdot,\ell\rangle_{1/2,\partial K}$ denotes the action of a linear functional ℓ in $H^{-1/2}(\partial K)$. For concise notation that reflects the element by element calculations, we use

$$(r,s)_{\Omega_h} := \sum_{K\in\Omega_h}(r,s)_K, \qquad \langle w,\ell\rangle_{\partial\Omega_h} := \sum_{K\in\Omega_h}\langle w,\ell\rangle_{1/2,\partial K}.$$

Note that in the latter definition, complex conjugations are absent, so to match conjugate linearity of other terms, we will often use notations like $\langle w, \bar{\ell}\rangle_{\partial\Omega_h}$ and $\langle \bar{w}, \ell\rangle_{\partial\Omega_h}$, whose meanings are self-explanatory. With these notations, the equations of the method derived from the integration by parts can be stated as follows:

$$\hat{\imath}\omega(\vec{u}\,,\vec{v}\,)_{\Omega_h} - (\phi, \vec{\nabla}\cdot\vec{v}\,)_{\Omega_h} + \langle\hat{\phi}, \overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h} = 0 \qquad \forall \vec{v} \in H(\mathrm{div}, \Omega_h), \tag{12a}$$

$$\hat{\imath}\omega(\phi,\eta)_{\Omega_h} - (\vec{u}\,, \vec{\nabla}\eta)_{\Omega_h} + \langle\overline{\eta}, \hat{u}_n\rangle_{\partial\Omega_h} = (f,\eta)_{\Omega}, \qquad \forall \eta \in H^1(\Omega_h). \tag{12b}$$

Above and throughout, all derivatives are taken element by element unless otherwise mentioned, and the 'broken' spaces are defined by

$$H(\mathrm{div}, \Omega_h) = \{\vec{\tau}:\ \vec{\tau}|_K \in H(\mathrm{div}, K),\ \forall K \in \Omega_h\},$$
$$H^1(\Omega_h) = \{v:\ v|_K \in H^1(K),\ \forall K \in \Omega_h\}.$$

From (12), it is clear that there are four solution components, namely 'interior' variables $(\vec{u}\,,\phi) \in L^2(\Omega)^N \times L^2(\Omega)$, and the numerical trace and flux $(\hat{u}_n, \hat{\phi})$ which lies in an affine space Q_g, which we now define. Let

$$R_g \stackrel{\text{def}}{=} \{(\vec{z}\,,\mu) \in H(\mathrm{div},\Omega)\times H^1(\Omega) : (\vec{z}\cdot\vec{n} - \mu)|_{\partial\Omega} = g\},$$
$$Q_g \stackrel{\text{def}}{=} \{(\hat{z}_n, \hat{\mu}) : \exists (\vec{z}\,,\mu) \in R_g \text{ such that } (\hat{z}_n,\hat{\mu}) = \mathrm{tr}_{\partial\Omega_h}(\vec{z}\,,\mu)\},$$

where $(\hat{z}_n,\hat{\mu}) = \mathrm{tr}_{\partial\Omega_h}(\vec{z}\,,\mu)$ signifies that for every mesh element $K \in \Omega_h$, we have

$$\hat{z}_n|_{\partial K} = \vec{z}\cdot\vec{n}|_{\partial K} \quad \text{ and } \quad \hat{\mu}|_{\partial K} = \mu|_{\partial K}.$$

In the case when $g = 0$, we simply denote $R = R_0$ and $Q = Q_0$. Note that the boundary condition (11c) becomes an *essential boundary condition* imposed in the numerical trace space Q_g. Observe that functions in Q_g, when restricted to the boundary of a single element ∂K, are in $H^{-1/2}(\partial K)\times H^{1/2}(\partial K)$. As already mentioned, the terms involving the numerical trace $\hat{\phi}$ and the numerical flux $\hat{u}_n$ in (12), are to be interpreted as $H^{-1/2}(\partial K)$-functional actions.

Let $(\vec{z}_g, \mu_g)$ in R_g and let $(\hat{z}_{g,n}, \hat{\mu}_g)$ be its corresponding trace in Q_g. We look for the solution, decomposed into

$$(\vec{u}\,,\phi,\hat{u}_n,\hat{\phi}) = (\vec{w},\varphi,\hat{w}_n,\hat{\varphi}) + (\vec{z}_g,\mu_g,\hat{z}_{g,n},\hat{\mu}_g).$$

The component $(\vec{z}_g,\mu_g,\hat{z}_{g,n},\hat{\mu}_g)$, consisting of the data and its extension, is known. Hence we only need to compute the unknown $(\vec{w},\varphi,\hat{w}_n,\hat{\varphi})$. Note that $(\hat{w}_n,\hat{\varphi})$ has homogeneous boundary conditions, i.e., it is in Q. We can compute an approximation to the unknown $(\vec{w},\varphi,\hat{w}_n,\hat{\varphi})$ by following the abstract program in § 2.1–§ 2.2, with these choices of spaces and forms:

$$\begin{aligned} b\big(\ (\vec{w},\varphi,\hat{w}_n,\hat{\varphi}),\ (\vec{v}\,,\eta)\ \big) :=& \hat{\imath}\omega(\vec{w},\vec{v}\,)_{\Omega_h} - (\varphi, \vec{\nabla}\cdot\vec{v}\,)_{\Omega_h} + \langle\hat{\varphi}, \overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h} \\ &+ \hat{\imath}\omega(\varphi,\eta)_{\Omega_h} - (\vec{w}, \vec{\nabla}\eta)_{\Omega_h} + \langle\overline{\eta}, \hat{w}_n\rangle_{\partial\Omega_h}, \end{aligned} \tag{13a}$$

$$l(\ (\vec{v}\,,\eta)\ ) := (f,\eta)_{\Omega} - (\hat{\imath}\omega\vec{z}_g + \nabla\mu_g, \vec{v}\,)_{\Omega} - (\hat{\imath}\omega\mu_g + \vec{\nabla}\cdot\vec{z}_g, \eta)_{\Omega}, \tag{13b}$$

$$U := L^2(\Omega)^N \times L^2(\Omega) \times Q, \tag{13c}$$

$$V := H(\mathrm{div}, \Omega_h) \times H^1(\Omega_h). \tag{13d}$$

The norm on V is defined by

$$\|(\vec{v}\,,\eta)\|_V^2 = \|\vec{\nabla}\eta + \hat{\imath}\omega\vec{v}\,\|_{\Omega_h}^2 + \|\hat{\imath}\omega\eta + \vec{\nabla}\cdot\vec{v}\,\|_{\Omega_h}^2 + \|\eta\|_{\Omega}^2 + \|\vec{v}\,\|_{\Omega}^2,$$

where the derivatives are calculated element by element as usual, while in contrast, the norm on R is defined using the global distributional derivatives:

$$\|(\vec{z},\mu)\|_R^2 = \|\vec{\nabla}\mu + \hat{\imath}\omega\vec{z}\|_\Omega^2 + \|\hat{\imath}\omega\mu + \vec{\nabla}\cdot\vec{z}\|_\Omega^2 + \|\vec{z}\|_\Omega^2 + \|\mu\|_\Omega^2.$$

This in turn defines the norm on Q by standard quotient topology, namely

$$\|(\hat{z}_n,\hat{\mu})\|_Q = \inf\left\{\|(\vec{z},\mu)\|_R \ : \ \forall(\vec{z},\mu)\in R \text{ such that } \mathrm{tr}_{\partial\Omega_h}(\vec{z},\mu) = (\hat{z}_n,\hat{\mu})\right\}. \tag{14}$$

The U-norm is then inherited from the product topology, i.e.,

$$\|(\vec{w},\varphi,\hat{w}_n,\hat{\varphi})\|_U^2 = \|\vec{w}\|_\Omega^2 + \|\varphi\|_\Omega^2 + \|(\hat{w}_n,\hat{\varphi})\|_Q^2. \tag{15}$$

Functions in Q are single valued on element interfaces by definition. They couple unknown interior values across the mesh elements.

3. The main result and discussion

Our main result is a wavenumber independent quasi-optimality estimate. In this section we state this result (Theorem 3.1). Its proof follows from two results proved in the next section. The remaining larger part of this section is devoted to a discussion on convergence rates and how the method compares with a number of recent works by other authors.

3.1. **Wavenumber independent quasi-optimality.** Our analysis is based on the following assumption.

Assumption 3. If ϕ satisfies

$$\Delta\phi + \omega^2\phi = F \qquad \text{on } \Omega \tag{16a}$$

$$\frac{\partial\phi}{\partial n} \pm \hat{\imath}\omega\phi = G \qquad \text{on } \partial\Omega \tag{16b}$$

for some $F\in L^2(\Omega)$ and $G\in H^{\frac{1}{2}}(\partial\Omega)$, then there is a $C>0$ (depending only on Ω) and an $\omega_0>0$ such that for all $\omega>\omega_0$, we have

$$\|\vec{\nabla}\phi\|_\Omega^2 + \omega^2\|\phi\|_\Omega^2 \le C\left(\|F\|_\Omega^2 + \|G\|_{\partial\Omega}^2\right). \tag{17}$$

This assumption is known to hold on bounded convex domains (see [21, Proposition 8.1.4]). It may hold more generally. In fact, a more general assumption of milder polynomial growth in the solution norm bound is assumed in [22, Assumption 4.7]. However, to our knowledge, while Assumption 3 has been verified for several cases, it is still a subject of active research to verify the more general forms of this assumption on specific domains.

Theorem 3.1: *Suppose Assumption 3 holds. Let $\mathcal{U} = (\vec{w},\varphi,\hat{w}_n,\hat{\varphi})\in U$ be the solution of the variational problem associated with the spaces and forms defined in* (13) *and let $\mathcal{U}_h$ denote its DPG approximation. Then there exist constants $\omega_0>0$ and $C>0$ such that the associated DPG solution $\mathcal{U}_h\in U_h$ satisfies the quasi-optimality estimate*

$$\|\mathcal{U}-\mathcal{U}_h\|_U \le C\inf_{\mathcal{W}_h\in U_h}\|\mathcal{U}-\mathcal{W}_h\|_U, \qquad \forall\omega>\omega_0.$$

Here, the constant C is independent of the wavenumber ω, the mesh Ω_h, and the approximating subspace U_h. The norm $\|\cdot\|_U$ is as in (15).

Proof. This follows from Theorem 2.1, whose assumptions –namely Assumptions 1 and 2– are verified in the next section (see Theorem 4.5 and Lemma 4.1). □

Note that so far, we have assumed nothing about the mesh Ω_h or the subspace $U_h \subseteq U$ in the above theorem. In fact, the theorem applies to arbitrary element shapes and any approximating subspace U_h built on any given Ω_h. However, it will be useful to consider a specific example of U_h obtained using a tetrahedral mesh to facilitate comparison with other works. We do this next.

3.2. **Tetrahedral convergence rates.** Let us now consider how to use Theorem 3.1 to obtain convergence rates when Ω_h is a geometrically conforming, shape regular, tetrahedral finite element mesh. Let $P_p(D)$ denote the set of functions that are restrictions of (multivariate) polynomials of degree at most p on a domain D. Let $p \geq 0$ and let

$$\begin{aligned} S_{h,p} &= \{\vec{w} : \vec{w}|_K \in P_p(K)^N\}, \qquad W_{h,p} = \{v : v|_K \in P_p(K)\}, \\ Q_{h,p} &= \big\{(\hat{z}_{n,h}, \hat{\mu}_h) : \quad \exists (\vec{z}_{h,p}, \mu_{h,p}) \in R \cap (S_{h,p} \times W_{h,p}) \\ &\qquad\qquad \text{such that } \mathrm{tr}_{\partial\Omega_h}(\vec{z}_{h,p}, \mu_{h,p}) = (\hat{z}_{n,h}, \hat{\mu}_h)\big\}. \end{aligned}$$

The example we want to consider is the case when the trial space is set by

$$U_h = S_{h,p} \times W_{h,p} \times Q_{h,p+1}. \tag{18}$$

Clearly, this is a subspace of the space U defined in (13c). We want to derive h and p convergence rates from Theorem 3.1. As usual, h denotes the maximum of the diameters of all mesh elements. We begin with the simplest case.

3.2.1. *The lowest order method.* We set $p = 0$ in (18) to get the lowest order method, i.e., the interior variables are approximated by piecewise constant approximations, while the numerical fluxes and traces are piecewise linear. We only need to study the rate at which the best approximation term in Theorem 3.1 converges.

To this end, let I_h denote the nodal interpolant of the *linear* Lagrange finite element, i.e., for a smooth function ψ, the interpolant $I_h\psi$ on any $K \in \Omega_h$ is the linear function whose values at the vertices of K equal the values of ψ there. By an abuse of notation, we use the same notation for vector functions, i.e., the vector function obtained by applying I_h to each component of $\vec{v}$ is denoted by $I_h\vec{v}$.

Let $(\vec{u}\,, \phi)$ solve the Helmholtz system (11). We tacitly assume that this pair is regular enough to apply the interpolant I_h. Since $(\vec{u}\,, \phi)$ satisfies the boundary condition of R, the approximation pair $(I_h\vec{u}\,, I_h\phi)$, by construction, is also in R (as can be seen by comparing the values of $I_h\vec{u} \cdot \vec{n}$ and ϕ at the Lagrange nodes on each face of K). Furthermore, $\mathrm{tr}_{\partial\Omega_h}(I_h\vec{u}\,, I_h\phi)$ is in $Q_{h,p+1}$, so

$$\inf_{(\hat{v}_{n,h}, \hat{\psi}_h) \in Q_{h,p+1}} \|(\hat{u}_n, \hat{\phi}) - (\hat{v}_{n,h}, \hat{\psi}_h)\|_Q \leq \|(\hat{u}_n, \hat{\phi}) - \mathrm{tr}_{\partial\Omega_h}(I_h\vec{u}\,, I_h\phi)\|_Q.$$

Since $\mathrm{tr}_{\partial\Omega_h}(\vec{u}\,, \phi) = (\hat{u}_n, \hat{\phi})$, by the definition of the Q-norm in (14), we have

$$\inf_{(\hat{v}_{n,h}, \hat{\psi}_h) \in Q_{h,p+1}} \|(\hat{u}_n, \hat{\phi}) - (\hat{v}_{n,h}, \hat{\psi}_h)\|_Q \leq \|(\vec{u}\,, \phi) - (I_h\vec{u}\,, I_h\phi)\|_R. \tag{19}$$

This is how we bound the best approximation terms for the numerical fluxes. The other terms forming the total best approximation error in Theorem 3.1 are easier. Combining

them with (19),

$$\|(u,\phi,\hat{\phi},\hat{u}_n)-(u_h,\phi_h,\hat{\phi}_h,\hat{u}_{n,h})\|_U^2 \le C\Big(\|(\vec{u}\,,\phi)-(I_h\vec{u}\,,I_h\phi)\|_R^2 + \inf_{\vec{w}_h\in S_{h,0}}\|\vec{u}-\vec{w}_h\|_\Omega^2 + \inf_{\psi_h\in W_{h,0}}\|\phi-\psi_h\|_\Omega^2\Big).$$

By the definition of the R-norm, this implies

$$\begin{aligned}\|(u,\phi,\hat{\phi},\hat{u}_n)-(u_h,\phi_h,\hat{\phi}_h,\hat{u}_{n,h})\|_U^2 \\ \le C\ \Big(\ \ \omega^2\|\vec{u}-I_h\vec{u}\,\|_\Omega^2 &+ \omega^2\|\phi-I_h\phi\|_\Omega^2 \\ +\ \|\vec{\nabla}\cdot(\vec{u}-I_h\vec{u}\,)\|_\Omega^2 &+ \|\vec{\nabla}(\phi-I_h\phi)\|_\Omega^2 \\ +\ \ \|\vec{u}-\Pi_S^0\vec{u}\,\|_\Omega^2 &+ \|\phi-\Pi_W^0\phi\|_\Omega^2\ \ \Big),\end{aligned} \tag{20}$$

where Π_S^0 and Π_W^0 denote the L^2-orthogonal projections into $S_{h,0}$ and $W_{h,0}$, resp. In the above two inequalities and throughout the paper we use C to denote a generic constant *independent of* ω. Its value may differ at different occurrences.

Convergence rates can now be concluded from (20). Note that we used flux and trace spaces of one higher order than the interior trial variables. This means that the middle two terms on the right hand side in (20) converges at the same h-rate as the last two terms. By using standard estimates for the L^2-projection and the nodal interpolant, (20) implies that

$$\begin{aligned}C\|(u,\phi,\hat{\phi},\hat{u}_n)-(u_h,\phi_h,\hat{\phi}_h,\hat{u}_{n,h})\|_U^2 \ \le\ & \omega^2h^2|\vec{u}\,|_{H^1(\Omega)}^2 + \omega^2h^2|\phi|_{H^1(\Omega)}^2 \\ +\ & h^2|\vec{u}\,|_{H^2(\Omega)}^2 + h^2|\phi|_{H^2(\Omega)}^2 \\ +\ & h^2|\vec{u}\,|_{H^1(\Omega)}^2 + h^2|\phi|_{H^1(\Omega)}^2.\end{aligned}$$

At this stage it is convenient to introduce a standard ω-dependent norm (see, e.g. [16]),

$$\|\phi\|_{s,\omega,D}^2 = \sum_{j=0}^{s}\omega^{2(s-j)}|\phi|_{H^j(D)}^2. \tag{21}$$

Note that if ϕ is a plane wave $e^{\hat{\imath}\omega x_\ell}$, then all the terms defining the norm scale with ω in the same way, namely as ω^{2s}. Hence, $\|\cdot\|_{s,\omega,D}$ is often considered a natural norm to use for wave propagation problems. We use this norm to summarize the conclusion of the above discussion.

COROLLARY 3.2: *The DPG solutions in the lowest order tetrahedral case satisfy*

$$\|\vec{u}-\vec{u}_h\|_\Omega + \|\phi-\phi_h\|_\Omega \le Ch\left(\|\phi\|_{2,\omega,\Omega}+\|\vec{u}\,\|_{2,\omega,\Omega}\right), \tag{22a}$$

$$\|(\hat{u}_n,\phi)-(\hat{u}_{n,h},\hat{\phi}_h)\|_Q \le Ch\left(\|\phi\|_{2,\omega,\Omega}+\|\vec{u}\,\|_{2,\omega,\Omega}\right). \tag{22b}$$

Remark 3.1. Note that although the solution component $\vec{u}$ is in $H(\mathrm{div},\Omega)$, we used the nodal $H^1(\Omega)$-interpolant to approximate it. We did so only because this is an easy way to find an approximating pair $(I_h\vec{u}\,,I_h\phi)$ that satisfies the Robin boundary condition of R. If one can find a more natural interpolant in R, one may be able to improve the regularity requirements of the above estimate (e.g., replace $\|u\|_{2,\omega,\Omega}$ by the more appropriate norm $\|\vec{\nabla}\cdot u\|_{1,\omega,\Omega}$).

Remark 3.2. A typical solution of the Helmholtz equation is the plane wave $\phi = e^{-\hat{\imath}\omega\vec{d}\cdot\vec{x}}$ and $\vec{u} = \phi\,\vec{d}$ for some unit vector $\vec{d}$ giving the direction of propagation. Then, (22) implies that

$$\|\vec{u} - \vec{u}_h\|_\Omega + \|\phi - \phi_h\|_\Omega \leq Ch\omega^2. \tag{23}$$

This estimate shows that *even if ωh is held constant, the errors increase with ω* for fixed data norms.

Once we fix ωh to be a constant, we may expect the relative best approximation error to remain more or less constant. Yet the discretization error may not. (Indeed, for the example in Remark 3.2, as we increase ω, even if we adjust h so that ωh remains constant, the discretization errors may grow with ω.) To our knowledge there is no finite element method to date that can provably avoid this problem in multiple space dimensions. A manifestation of this error increase with ω, for the standard methods, is via the accumulation of *phase errors* in the direction of wave propagation. This was expounded in [2] in the context of the standard method for the Helmholtz equation, but it also well recognized for Maxwell (see e.g. [14, Fig. 6]) and other wave phenomena.

Surprisingly however, for the DPG method, phase errors were *observed to be negligible* in all our numerical experiments. (See Section 5 for an extended discussion.) We are currently unable to explain this superior performance theoretically. It is possible that the error bounds of Corollary 3.2 are too pessimistic.

3.2.2. *Higher order convergence.* Next, consider the case $p \geq 1$. Both h and p convergence rates can be derived from Theorem 3.1 because the constant in the theorem is independent of p. We will now need a conforming p-optimal $H^1(\Omega)$-interpolant, e.g., the one given in [11] and [10, Theorem 8.1], that satisfies

$$\|\psi - \Pi_{hp}\psi\|_\Omega + h\|\vec{\nabla}(\psi - \Pi_{hp}\psi)\|_\Omega \leq C\ln(\tilde{p})^2\, h^{s+1}\tilde{p}^{-s}|\psi|_{H^{s+1}(\Omega)},$$

for all ψ in $H^{s+1}(\Omega)$ with $3/2 < s+1 \leq p+1$. Above, $\tilde{p} = \max(p, 2)$. For vector functions $\vec{v}$, let $\Pi_{hp}\vec{v}$ denote the vector function obtained by applying Π_{hp} to each component of $\vec{v}$. By following the construction of Π_{hp} (see [11]) it is easy to see that the approximating pair $(\Pi_{hp}\vec{u}, \Pi_{hp}\phi)$ satisfies the boundary condition of R whenever the pair $(\vec{u}, \phi)$ is in R. Now, we can proceed as in the lowest order case (cf. (20)) to obtain

$$\begin{aligned}
\|\vec{u} - \vec{u}_h\|_\Omega + \|\phi - \phi_h\|_\Omega + \|(\hat{u}_n, \phi) - (\hat{u}_{n,h}, \hat{\phi}_h)\|_Q \\
\leq C \Big(& \omega\ln(\tilde{p})^2 h^s\tilde{p}^{-s}|\vec{u}\,|_{H^s(\Omega)} + \omega\ln(\tilde{p})^2 h^s\tilde{p}^{-s}|\phi|_{H^s(\Omega)} \\
& + \ln(\tilde{p})^2 h^s\tilde{p}^{-s}|\vec{u}\,|_{H^{s+1}(\Omega)} + \ln(\tilde{p})^2 h^s\tilde{p}^{-s}|\phi|_{H^{s+1}(\Omega)} \\
& + h^s\tilde{p}^{-s}|\vec{u}\,|_{H^s(\Omega)} + h^s\tilde{p}^{-s}|\phi|_{H^s(\Omega)} \Big).
\end{aligned}$$

Overestimating, we can immediately summarize an estimate for the interior variables using the norm defined in (21).

COROLLARY 3.3: *The DPG solution in the higher order tetrahedral case satisfies*

$$\|\vec{u} - \vec{u}_h\|_\Omega + \|\phi - \phi_h\|_\Omega \leq Ch^s\frac{\ln(\tilde{p})^2}{\tilde{p}^s}\left(\|\phi\|_{s+1,\omega,\Omega} + \|\vec{u}\,\|_{s+1,\omega,\Omega}\right), \tag{24}$$

for all $s = 1, 2, \ldots, p$.

Again, we emphasize that C is independent of ω, h, and p. Obviously a similar estimate can also be stated for the numerical traces and fluxes.

3.2.3. *Comparison with other recent works.* Next, we want to compare the above stated convergence rates with a few other recent works that state error estimates explicitly showing the wavenumber dependence. Note that our method simultaneously gives error estimates for both the pressure ϕ and velocity $\vec{u}$. Most other methods give only an approximation to the primal variable ϕ. An approximation to the velocity variable $\vec{u}$ must then be derived by numerical differentiation (but this results in a loss of convergence order for those methods). Hence, we will only compare error estimates for ϕ.

(A) *The standard p FEM:* New estimates for this old method have been derived recently in [22]. They show that for domains with analytic boundary, or on convex polygons, if

$$\frac{\omega h}{p} \text{ is small} \qquad \text{and} \qquad p \geq C \log \omega, \tag{25}$$

and additionally if the Helmholtz solution operator's norm satisfies a polynomial growth assumption (satisfied if Assumption 3 holds) then the solution ϕ_h of the standard p-finite element method satisfies

$$\omega\|\phi-\phi_h\|_\Omega + \|\vec{\nabla}(\phi-\phi_h)\|_{\Omega_h} \leq C \inf_{\psi_h \in W_{h,p}\cap H^1(\Omega)} \Big(\omega\|\phi-\psi_h\|_\Omega + \|\vec{\nabla}(\phi-\psi_h)\|_{\Omega_h}\Big)$$

with a C independent of ω. This is perhaps the clearest precise statement available in the literature demonstrating that pollution effects are removed in high order p FEM. This estimate is better than the estimate of our Theorem 3.1. Yet, the numerical performance of our method (in the case of low p, as reported later) is better than the standard FEM. The main advantage in our theory is that we have no need for condition (25). The DPG method has better pre-asymptotic stability properties (e.g., we have no need to assume a sufficiently small h) and yields Hermitian positive definite matrices. The growth of conditioning with h of both methods are similar.

(B) *The plane wave DG method (PWDG):* In two space dimensions, the recent paper [16, Theorem 3.14] analyzes a Trefftz DG method using plane waves for trial subspaces. If plane waves in $p' = 2m+1$ wave directions (sufficiently separated) are used with each mesh element, then *for sufficiently large* p', they prove that

$$\omega\|\phi-\phi_h\|_\Omega \leq C(\omega h)\,\mathrm{diam}(\Omega)\,h^{s-1}\left(\frac{\ln(p')}{p'}\right)^{s-1/2}\|\phi\|_{s+1,\omega,\Omega} \tag{26}$$

holds for all $s < \lceil (m+1)/2 \rceil$, where $C(\omega h)$ is an increasing function of ωh. For the sake of comparison when ωh is fixed, we may multiply both sides of (26) by h so that both (26) and our estimate (24) gives the same h-convergence rate. If one agrees to view our polynomial degree p to be more or less comparable to their parameter p', then our estimate is comparable to theirs (with the difference that we neither have ω dependence in C, nor do we need to assume large p). However, since we work with polynomial spaces, we avoid the conditioning problems they faced due to the use of plane waves.

We should note however that the numerical results reported in [16] are excellent. It begs the question if we could use their same plane wave basis functions to form the trial space U_h in our DPG setting. Indeed, this can be done. Note that in Theorem 3.1 we placed *no* assumptions on U_h, so the theorem applies verbatim in this setting. The only theoretical difficulty is in bounding the best approximation error estimate. While the best

approximation estimates for the interior variables immediately follow from [16], bounding the flux best approximation terms seems to require conforming plane wave approximation estimates not available in [16].

(C) *Interior penalty DG method with complex stabilization:* This method was recently developed in [12, 13]. In the lowest order case [12, Theorem 5.5 and Eq.(6.6)] they prove that

$$\|\vec{\nabla}(\phi-\phi_h)\|_{\Omega_h} \leq (1+\omega h)(C_1\omega h + C_2\omega^3 h^2)$$

for a specific choice of their stabilization parameters. Here C_1 and C_2 depend only on the load and are independent of ω and h. We may compare this to our estimate (23). If ωh is held fixed, then the growth with respect to ω in both the estimates are linear. In [13], the higher order case, for a general polynomial degree p, is considered. The best estimate they have, after an iterative improvement [13, Theorem 5.1], is

$$\|\phi-\phi_h\|_{\Omega} \leq C\frac{\omega h^{\min(p+1,s)}}{p^s}\|\phi\|_{H^s(\Omega)}, \tag{27}$$

provided

$$\frac{\omega^3 h^2}{p} \leq C. \tag{28}$$

The estimate (27) is comparable to (24). A notable difference is the absence of factors of $\ln(p)$ in their estimate: These factors arose due to our need to use conforming p-optimal projectors, a need absent in traditional DG analyses. Also note that we have no need for assumption (28).

4. Analysis

In this section, we verify the assumptions of Theorem 2.1 for the DPG method applied to the Helmholtz equation. Throughout this section, we tacitly assume that Ω is such that Assumption 3 holds for Helmholtz solutions. We show that Assumption 3 implies Assumptions 1 and 2 for the Helmholtz application.

4.1. **Verification of injectivity.**

To prove the following lemma we only use a weak consequence of Assumption 3, namely that (17) implies uniqueness of solutions for the Helmholtz problem (16).

Lemma 4.1: *Assumption 1 holds for the DPG sequilinear form defined in* (13).

Proof. Consider any $(\vec{u}\,,\phi,\hat{u}_n,\hat{\phi}) \in U = L^2(\Omega)^N \times L^2(\Omega) \times Q$ satisfying

$$\begin{aligned} \hat{\imath}\omega(\vec{u}\,,\vec{v}\,)_{\Omega_h} - (\phi,\vec{\nabla}\cdot\vec{v}\,)_{\Omega_h} + \langle\hat{\phi},\overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h} = 0,\\ \hat{\imath}\omega(\phi,\eta)_{\Omega_h} - (\vec{u}\,,\vec{\nabla}\eta)_{\Omega_h} + \langle\overline{\eta},\hat{u}_n\rangle_{\partial\Omega_h} = 0, \end{aligned} \tag{29}$$

for all $(\vec{v},\eta) \in V = H(\mathrm{div},\Omega_h)\times H^1(\Omega_h)$. Testing with functions in the subspace of V consisting of globally infinitely differentiable functions, compactly supported on Ω, we find that

$$\hat{\imath}\omega\vec{u} + \nabla\phi = 0 \quad \text{and} \quad \hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{u} = 0 \tag{30}$$

in the sense of distributions on the open set Ω. This implies that $\phi \in H^1(\Omega)$ and $\vec{u} \in H(\mathrm{div},\Omega)$, which now allows us to integrate by parts in (29). Hence, for every

$(\vec{v},\eta)\in V$, we obtain the equations

$$\left\langle \hat{\phi}-\phi, \overline{\vec{v}\cdot\vec{n}}\right\rangle_{\partial\Omega_h}=0 \quad \text{and} \quad \langle\overline{\eta},\hat{u}_n-\vec{u}\cdot\vec{n}\rangle_{\partial\Omega_h}=0.$$

Thus, $(\hat{u}_n,\hat{\phi})=\mathrm{tr}_{\partial\Omega_h}(\vec{u}\,,\phi)$. Furthermore, since $(\hat{u}_n,\hat{\phi})\in Q$, we satisfy the boundary condition

$$\vec{u}\cdot\vec{n}-\phi=0, \qquad \text{over } \partial\Omega. \tag{31}$$

Now, we test equation (29) with the globally conforming $\eta=\phi\in H^1(\Omega)$ to get

$$\hat{\imath}\omega\,\|\phi\|_{\Omega}^2-(\vec{u},\nabla\phi)_{\Omega}+\|\phi\|_{\partial\Omega}^2=0.$$

Since $\vec{\nabla}\phi=-\hat{\imath}\omega\vec{u}$ by (30), the real part of this equation implies that $\|\phi\|_{\partial\Omega}^2=0$. Hence, by (31), $\phi=\vec{u}\cdot\vec{n}=0$ on $\partial\Omega$. Thus, ϕ satisfies, in a distributional sense, the Helmholtz boundary value problem (16) with zero F and G. Therefore, by Assumption 3, $\phi=0$ in Ω. Then, we obviously also have $\vec{u}=\vec{0}$ in Ω, $\hat{\phi}|_{\partial K}=\phi|_{\partial K}=0$ and $\hat{u}_n|_{\partial K}=\vec{u}\cdot\vec{n}|_{\partial K}=0$ for all $K\in\Omega_h$. □

4.2. **Optimal test norm.** The optimal norm is easily calculated from its definition (4). Let $\mathcal{U}=(\vec{w},\varphi,\hat{w}_n,\hat{\varphi})\in U=L^2(\Omega)^N\times L^2(\Omega)\times Q$. Then the bilinear form in (13) can be written as

$$b\big(\,\mathcal{U}\,,(\vec{v}\,,\eta)\,\big)=-(\vec{w},\vec{\nabla}\eta+\hat{\imath}\omega\vec{v}\,)_{\Omega_h}-(\varphi,\hat{\imath}\omega\eta+\vec{\nabla}\cdot\vec{v}\,)_{\Omega_h}+\langle\hat{\varphi},\overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h}+\langle\overline{\eta},\hat{w}_n\rangle_{\partial\Omega_h}.$$

It is easy to check that in this case, the supremum in the optimal test norm equals

$$\|(\vec{v}\,,\eta)\|_{\mathrm{opt},V}^2=\|\vec{\nabla}\eta+\hat{\imath}\omega\vec{v}\,\|_{\Omega_h}^2+\|\hat{\imath}\omega\eta+\vec{\nabla}\cdot\vec{v}\,\|_{\Omega_h}^2+\big|[\vec{v}\,,\eta]\big|_{\partial\Omega_h}^2,$$

where

$$\big|[\vec{v}\,,\eta]\big|_{\partial\Omega_h}=\sup_{(\hat{w}_n,\hat{\varphi})\in Q}\frac{\big|\langle\overline{\eta},\hat{w}_n\rangle_{\partial\Omega_h}+\langle\hat{\varphi},\overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h}\big|}{\|(\hat{w}_n,\hat{\varphi})\|_Q}.$$

By the definition of the norm on Q, this can be rewritten as

$$\big|[\vec{v}\,,\eta]\big|_{\partial\Omega_h}=\sup_{(\vec{z},\mu)\in R}\frac{\big|\langle\overline{\eta},\vec{z}\cdot\vec{n}\rangle_{\partial\Omega_h}+\langle\mu,\overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h}\big|}{\|(\vec{z}\,,\mu)\|_R}.$$

4.3. **Norm Equivalence.** Now we turn our attention to verifying Assumption 2. The main result of this subsection is Theorem 4.5, which verifies the assumption. We begin with the following lemma. Recall that the generic constant C is independent of ω throughout.

LEMMA 4.2: *Given any η in $L^2(\Omega)$, there is an $\vec{r}\in H(\mathrm{div},\Omega)$ and a $\phi\in H^1(\Omega)$ such that*

$$\hat{\imath}\omega\vec{r}+\vec{\nabla}\phi=0, \qquad \text{on } \Omega \tag{32a}$$

$$\hat{\imath}\omega\phi+\vec{\nabla}\cdot\vec{r}=\eta, \qquad \text{on } \Omega \tag{32b}$$

$$\vec{r}\cdot\vec{n}=\pm\phi \qquad \text{on } \partial\Omega, \tag{32c}$$

and

$$\|(\vec{r}\,,\phi)\|_R\le C\|\eta\|_{\Omega}. \tag{33}$$

Proof. Define the sesquilinear form

$$a_\pm(\varphi,\psi) \stackrel{\text{def}}{=} (\vec{\nabla}\varphi,\vec{\nabla}\psi)_\Omega - \omega^2(\varphi,\psi)_\Omega \pm \hat{\imath}\omega\langle\varphi,\psi\rangle_{\partial\Omega}.$$

Let $\phi \in H^1(\Omega)$ be the (unique) solution of

$$a_\pm(\phi,\psi) = (\hat{\imath}\omega\eta,\psi)_\Omega, \qquad \forall\psi \in H^1(\Omega).$$

Then, set $\vec{r}$ by $\hat{\imath}\omega\vec{r} = -\vec{\nabla}\phi$. It is easy to see that $\vec{r} \in H(\mathrm{div},\Omega)$ and (32) is satisfied. To prove (33), we note that by Assumption 3, we have $\|\vec{\nabla}\phi\|_\Omega^2 + \omega^2\|\phi\|_\Omega^2 \le C\|\hat{\imath}\omega\eta\|_\Omega^2$. Since $\|\hat{\imath}\omega\vec{r} + \vec{\nabla}\phi\|_{\Omega_h} = 0$, $\|\hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{r}\|_{\Omega_h} = \|\eta\|_\Omega$ and $\omega\|\vec{r}\|_\Omega = \|\vec{\nabla}\phi\|_\Omega$, we immediately have

$$\omega^2\|(\vec{r},\phi)\|_R^2 = \omega^2(\|\hat{\imath}\omega\vec{r} + \vec{\nabla}\phi\|_{\Omega_h}^2 + \|\hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{r}\|_{\Omega_h}^2 + \|\vec{r}\|_\Omega^2 + \|\phi\|_\Omega^2) \le C\omega^2\|\eta\|_\Omega^2,$$

which is (33). □

The next lemma is the analogue of Lemma 4.2 for the vector test variable $\vec{v} \in L^2(\Omega)^N$.

LEMMA 4.3: *There exists an $\omega_1 > 0$ such that given any $\vec{v}$ in $L^2(\Omega)^N$, there is an $\vec{r} \in H(\mathrm{div},\Omega)$ and a $\phi \in H^1(\Omega)$ satisfying*

$$\hat{\imath}\omega\vec{r} + \vec{\nabla}\phi = \vec{v}\,, \qquad \text{on } \Omega \tag{34a}$$

$$\hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{r} = 0, \qquad \text{on } \Omega \tag{34b}$$

$$\vec{r}\cdot\vec{n} = \pm\phi \qquad \text{on } \partial\Omega, \tag{34c}$$

and

$$\|(\vec{r},\phi)\|_R \le C\|\vec{v}\,\|_\Omega \tag{35}$$

for all $\omega > \omega_1$.

Proof. First, we set $\phi \in H^1(\Omega)$ to be the (unique) solution of

$$a_\pm(\phi,\psi) = (\vec{v}\,,\vec{\nabla}\psi)_\Omega, \qquad \forall\psi \in H^1(\Omega). \tag{36}$$

Then, set $\vec{r}$ by $\hat{\imath}\omega\vec{r} = -\vec{\nabla}\phi + \vec{v}$. It is easy to see that $\vec{r} \in H(\mathrm{div},\Omega)$ and (34) is satisfied. It only remains to prove (35). For this, we pick ψ in (36) as $\psi = \phi + \zeta$ where $\zeta \in H^1(\Omega)$ is the unique solution of the adjoint problem:

$$a_\pm(\varphi,\zeta) = 2\omega^2(\varphi,\phi)_\Omega \qquad \forall\varphi \in H^1(\Omega). \tag{37}$$

By Assumption 3, we clearly have:

$$\|\vec{\nabla}\zeta\|_\Omega^2 + \omega^2\|\zeta\|_\Omega^2 \le C\omega^4\|\phi\|_\Omega^2. \tag{38}$$

Moreover, by (37),

$$\begin{aligned}\mathrm{Re}\big(a_\pm(\phi,\phi+\zeta)\big) &= \mathrm{Re}\big(a_\pm(\phi,\phi)\big) + \mathrm{Re}\big(a_\pm(\phi,\zeta)\big)\\ &= \|\vec{\nabla}\phi\|_\Omega^2 - \omega^2\|\phi\|_\Omega^2 + \mathrm{Re}(a(\phi,\zeta))\\ &= \|\vec{\nabla}\phi\|_\Omega^2 + \omega^2\|\phi\|_\Omega^2.\end{aligned}$$

Hence,

$$\begin{aligned}\|\vec{\nabla}\phi\|_\Omega^2+\omega^2\|\phi\|_\Omega^2 &= \mathrm{Re}\big(a_\pm(\phi,\phi+\zeta)\big)=\big(\vec{v},\vec{\nabla}(\phi+\zeta)\big)_\Omega\\ &\le \|\vec{v}\,\|_\Omega\left(\|\vec{\nabla}\phi\|_\Omega+\|\vec{\nabla}\zeta\|_\Omega\right)\\ &\le C\|\vec{v}\,\|_\Omega\left(\|\vec{\nabla}\phi\|_\Omega^2+\omega^4\|\phi\|_\Omega^2\right)^{1/2}.\\ &\le C\,\omega\|\vec{v}\,\|_\Omega\left(\frac{1}{\omega^2}\|\vec{\nabla}\phi\|_\Omega^2+\omega^2\|\phi\|_\Omega^2\right)^{1/2}.\end{aligned}$$

Thus, for large ω, we obtain $\|\vec{\nabla}\phi\|_\Omega^2+\omega^2\|\phi\|_\Omega^2\le C\omega^2\|\vec{v}\,\|_\Omega^2$. Using also the equalities $\|\hat{\imath}\omega\vec{r}+\vec{\nabla}\phi\|_{\Omega_h}=\|\vec{v}\,\|_\Omega$, $\|\hat{\imath}\omega\phi+\vec{\nabla}\cdot\vec{r}\,\|_{\Omega_h}=0$, and $\omega\|\vec{r}\,\|_\Omega=\|\vec{v}-\vec{\nabla}\phi\|_\Omega$, we obtain

$$\begin{aligned}\|(\vec{r},\phi)\|_R^2 &= \|\hat{\imath}\omega\vec{r}+\vec{\nabla}\phi\|_{\Omega_h}^2+\|\hat{\imath}\omega\phi+\vec{\nabla}\cdot\vec{r}\,\|_{\Omega_h}^2+\|\vec{r}\,\|_\Omega^2+\|\phi\|_\Omega^2\\ &\le \|\vec{v}\,\|_\Omega^2+\frac{1}{\omega^2}\|\vec{v}-\vec{\nabla}\phi\|_\Omega^2+\|\phi\|_\Omega^2\\ &\le \left(1+\frac{2}{\omega^2}\right)\|\vec{v}\,\|_\Omega^2+\frac{C}{\omega^2}\left(\|\vec{\nabla}\phi\|_\Omega^2+\omega^2\|\phi\|_\Omega^2\right)\\ &\le C\left(1+\frac{1}{\omega^2}\right)\|\vec{v}\,\|_\Omega^2.\end{aligned}$$

Hence the estimate of the lemma follows taking ω large enough. □

LEMMA 4.4: *There is an $\omega_1>0$ such that for any $(\vec{v},\eta)\in V$ we have*

$$\|\vec{v}\,\|_\Omega+\|\eta\|_\Omega\le C\,\|(\vec{v},\eta)\|_{\mathrm{opt},V}\,,$$

for all $\omega>\omega_1$.

Proof. For a given $\vec{v}$ and η, apply Lemmas 4.2 and 4.3 to obtain $(\vec{r},\phi)\in R$ satisfying

$$\hat{\imath}\omega\vec{r}+\vec{\nabla}\phi=\vec{v}\,,\qquad\text{on }\Omega \tag{39a}$$

$$\hat{\imath}\omega\phi+\vec{\nabla}\cdot\vec{r}=\eta,\qquad\text{on }\Omega \tag{39b}$$

and

$$\|(\vec{r},\phi)\|_R\le C\left(\|\vec{v}\,\|_\Omega+\|\eta\|_\Omega\right). \tag{40}$$

Then

$$\begin{aligned}\|\vec{v}\,\|_\Omega^2+\|\eta\|_\Omega^2 &= (\hat{\imath}\omega\vec{r}+\vec{\nabla}\phi,\vec{v}\,)_{\Omega_h}+(\hat{\imath}\omega\phi+\vec{\nabla}\cdot\vec{r},\eta)_{\Omega_h}\qquad\text{(by (39)),}\\ &= -(\vec{r},\hat{\imath}\omega\vec{v}+\vec{\nabla}\eta)_\Omega+(\vec{\nabla}\phi,\vec{v}\,)_\Omega+(\vec{r},\vec{\nabla}\eta)_{\Omega_h}\\ &\quad-(\phi,\hat{\imath}\omega\eta+\vec{\nabla}\cdot\vec{v}\,)_{\Omega_h}+(\vec{\nabla}\cdot\vec{r},\eta)_{\Omega_h}+(\phi,\vec{\nabla}\cdot\vec{v}\,)_{\Omega_h}\\ &= -(\vec{r},\hat{\imath}\omega\vec{v}+\vec{\nabla}\eta)_\Omega-(\phi,\hat{\imath}\omega\eta+\vec{\nabla}\cdot\vec{v}\,)_{\Omega_h}\\ &\quad+\langle\phi,\overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h}+\langle\overline{\eta},\vec{r}\cdot\vec{n}\rangle_{\partial\Omega_h}\qquad\text{(integrating by parts).}\end{aligned}$$

By Cauchy-Schwarz inequality,

$$\|\vec{v}\,\|_{\Omega}^2 + \|\eta\|_{\Omega}^2 \leq \left(\|\vec{\nabla}\eta + \hat{\imath}\omega\vec{v}\,\|_{\Omega_h}^2 + \|\hat{\imath}\omega\eta + \vec{\nabla}\cdot\vec{v}\,\|_{\Omega_h}^2\right)^{1/2} \|(\vec{r},\phi)\|_R + \left(\frac{\left|\langle\phi, \overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h} + \langle\overline{\eta}, \vec{r}\cdot\vec{n}\rangle_{\partial\Omega_h}\right|}{\|(\vec{r},\phi)\|_R}\right) \|(\vec{r},\phi)\|_R.$$

so the result follows from (40). □

THEOREM 4.5: *Suppose Assumption 3 holds. Then, there are positive constants* ω_0, C_1, C_2 *such that for all* $\omega > \omega_0$ *and for all* $(\vec{v}\,,\eta) \in H(\mathrm{div},\Omega_h) \times H^1(\Omega_h)$,

$$C_1\|(\vec{v}\,,\eta)\|_V \leq \|(\vec{v}\,,\eta)\|_{\mathrm{opt},V} \leq C_2\|(\vec{v}\,,\eta)\|_V.$$

Proof.

Upper bound: We only need to bound the jump term (as the other terms are present in the V-norm). Since

$$\begin{aligned}\langle\mu, \overline{\vec{v}\cdot\vec{n}}\rangle_{\partial\Omega_h} + \langle\overline{\eta}, \vec{z}\cdot\vec{n}\rangle_{\partial\Omega_h} &= (\vec{\nabla}\mu, \vec{v}\,)_{\Omega_h} + (\mu, \vec{\nabla}\cdot\vec{v}\,)_{\Omega_h} + (\vec{z}\,, \vec{\nabla}\eta)_{\Omega_h} + (\vec{\nabla}\cdot\vec{z}\,, \eta)_{\Omega_h} \\ &= (\vec{\nabla}\mu + \hat{\imath}\omega\vec{z}\,, \vec{v}\,)_{\Omega_h} + (\mu, \hat{\imath}\omega\eta + \vec{\nabla}\cdot\vec{v}\,)_{\Omega_h} \\ &\quad + (\vec{z}\,, \vec{\nabla}\eta + \hat{\imath}\omega\vec{v}\,)_{\Omega_h} + (\hat{\imath}\omega\mu + \vec{\nabla}\cdot\vec{z}\,, \eta)_{\Omega_h} \\ &\leq \|(\vec{v}\,,\eta)\|_V\|(\vec{z}\,,\mu)\|_R,\end{aligned}$$

we can divide by $\|(\vec{z}\,,\mu)\|_R \neq 0$ and take the supremum on R to obtain the upper bound.

Lower bound: We only need to estimate the two terms in the V-norm that are not in the optimal norm. But these bounds are immediate by Lemma 4.4 provided ω is large enough. This finishes the proof. □

5. Numerical experiments

In this section we present results of numerical experiments for four distinct model problems. In all cases, the results are better than predicted by the preceding theory. Let us begin by first describing precisely the spaces of approximation used in our computations, and the discrete approximation of the operator T which generates the optimal test space.

We need to specify the trial space (cf. (13c) and (18)). This is constructed using a mesh Ω_h of *quadrilaterals*. Let $\mathcal{Q}^{(l)}$ denote the space of polynomials in one variable of degree at most l, and let $\mathcal{Q}^{(l,m)}$ denote the space of polynomials of degree at most l and m in the two variables x_1 and x_2, resp. We use it to define the sequence of spaces

$$X_p(\hat{K}) = \mathcal{Q}^{(p,p)}, \quad Y_p(\hat{K}) = \mathcal{Q}^{(p,p-1)} \times \mathcal{Q}^{(p-1,p)}, \quad X_{p-1}(\hat{K}) = \mathcal{Q}^{(p-1,p-1)},$$

for $p \geq 1$. These sequence of spaces are to be interpreted as subspaces of $H^1(\hat{K})$, $H(\mathrm{div},\hat{K})$, and $L^2(\hat{K})$, resp., where $\hat{K} = (0,1)^2$. The spaces $X_p(K)$, $Y_p(K)$, $X_{p-1}(K)$ on a general $K \in \Omega_h$ are the analogous spaces of shape functions on the physical element defined through the usual pullback (depending on the subspace interpretation) to the reference element. Similarly, let $\mathcal{Q}^{(l)}(E)$ on a (possibly curved) mesh edge E denote the image of $\mathcal{Q}^{(l)}$ from $(0,1)$ under the standard map. The interior variables are approximated in $S_h = \{\vec{r} : \vec{r}\,|_K \in X_p(K)^N\}$ and $W_h = \{v : v|_K \in X_p(K)\}$, resp., while the numerical fluxes and

traces lie in $Q_h = \big\{(\hat{z}_{n,h}, \hat{\mu}_h) \in R :\ \hat{z}_{n,h}|_E$ and $\hat{\mu}_h|_E$ are in $\mathcal{Q}^{(p+1)}(E)$ for all mesh edges E, and $\hat{z}_{n,h} - \hat{\mu}_h = 0$ on $\partial\Omega\big\}$, i.e.,

$$U_h = S_h \times W_h \times Q_h.$$

Note that the numerical traces (in the second component of Q_h) are continuous at edges that meet at a vertex.

As noted in § 2.3, the practical application of DPG requires us to approximate the operator T by a discrete version $\tilde{T}$ which maps into a finite dimensional enriched test space $\tilde{V} \subset V$. In our experiments, $\tilde{V}$ is constructed by considering the local polynomial order p of the element K and a global parameter Δp. We compute the practical test space using (10), with $\tilde{V}$ set element by element by

$$\tilde{V}|_K = Y_{p+\Delta p}(K) \times X_{p+\Delta p}(K) \subseteq H(\mathrm{div}, K) \times H^1(K).$$

In all our experiments, we fix $\Delta p = 2$. This choice is dictated by our previous computational experience [6, 8, 9], whereby it was clear that using a higher Δp did not result in any error improvements, while using a lower Δp could result in a non-injective $\tilde{T}$.

5.1. **Model problem A.** This is problem (11), for which we provided a theoretical analysis. Our domain Ω is the square $(0,1)^2$. The right hand sides f and g of (11b) and (11c), respectively, are set so that the exact solution is a plane wave solution (propagating in the θ-direction), namely

$$\phi(\vec{x}) = e^{-i\omega(x_1\cos\theta + x_2\sin\theta)}, \quad \vec{u}(\vec{x}) = \phi(\vec{x})(\cos\theta, \sin\theta),$$

and correspondingly,

$$f = 0, \quad g(\vec{x}) = \phi(\vec{x})[n_1(\vec{x})\cos\theta + n_2(\vec{x})\sin\theta - 1]$$

where the outward unit normal vector is $\vec{n} = (n_1, n_2)$.

We employ a grid of square, bilinear elements (with h denoting the length of their side) to discretize the problem, i.e., $p = 1$ in the above definition of U_h. Our mesh is quite coarse, with $\omega h = \pi/2$, equivalent to four elements per wavelength when the direction is aligned with the mesh. This is still enough to reasonably represent the wave, as the L^2 best approximation error (BAE) will vary from about 6.5% to 9%, depending on the direction θ (the case of diagonal propagation $\theta = \pi/4$, being the best case). We use BAE to denote the best approximation error in both $\vec{u}$ and ϕ, i.e.,

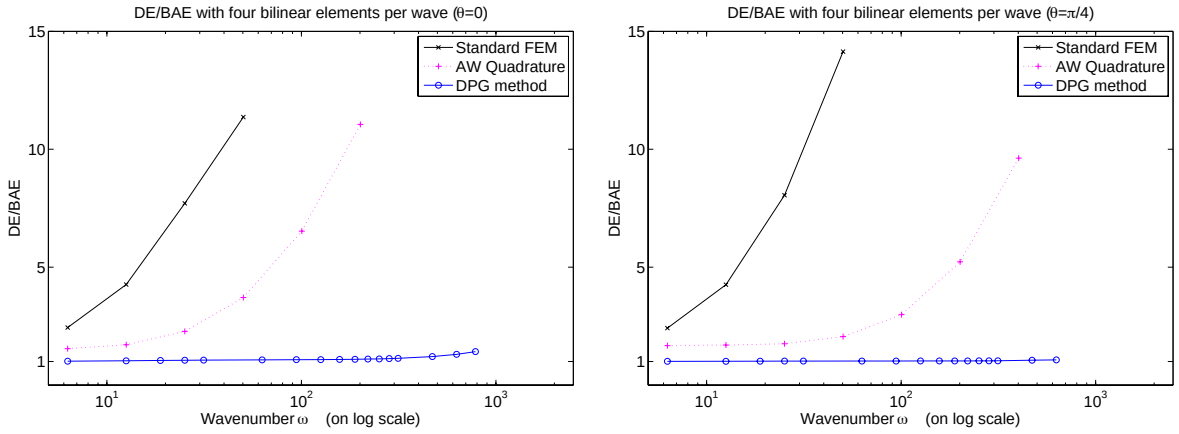
$$(\mathrm{BAE})^2 = \inf_{\vec{w}_h \in S_h} \|\vec{u} - \vec{w}_h\|_\Omega^2 + \inf_{\varphi_h \in W_h} \|\phi - \varphi_h\|_\Omega^2,$$

while DE denotes the discretization error defined by

$$(\mathrm{DE})^2 = \|\vec{u} - \vec{u}_h\|_\Omega^2 + \|\phi - \phi_h\|_\Omega^2.$$

We will report the ratio DE/BAE. Due to Theorem 3.1, we expect this ratio to be bounded by a mesh-independent constant. We are interested in how this ratio changes with ω. Since BAE represents the best any method can do using the given spaces, a study of how the ratio DE/BAE changes with respect to ω can give us an idea of how pollution effects influence discretization errors.

We compare the performance of the DPG method to two other methods readily available in our software package, namely the standard FEM, and the FEM using the new and interesting quadrature rules of [1]. Note that when considering these other methods, BAE is to be interpreted as the L^2 best approximation error of the space used by the other



(a) The $\theta = 0$ case (b) The $\theta = \pi/4$ case

FIGURE 1. Model A: The ratio of discretization errors to best approximation erros in $L^2(\Omega)$-norm for two plane wave directions.

method, i.e., BAE$= \inf_{\varphi_h \in H^1(\Omega) \cap W_h} \|\phi - \varphi_h\|_\Omega$, and similarly, DE denotes the discretization error of the other method. Note also that we use a very modest number of mesh points per wavelength (e.g., in comparison to results from other least square methods in the literature [20]).

The results are in Figure 1 for two values of θ. We observe that in both cases the quality of standard finite element approximations quickly deteriorates as we increase the wavenumber. The deterioration exists, but is delayed, when the method of [1] is used – see the curve labeled "AW Quadrature". The DPG solutions however are still very close to the best approximations.

5.2. **Model problem B.** This problem is similar to Model Problem A. We keep all parameters the same as before, but use slightly different boundary conditions. These boundary conditions are typically used to demonstrate the accumulation of phase errors in the direction of wave propagation for standard Galerkin methods. The boundary value problem is:

$$\hat{\imath}\omega\vec{u} + \vec{\nabla}\phi = \vec{0}, \qquad \text{on } \Omega \tag{41a}$$

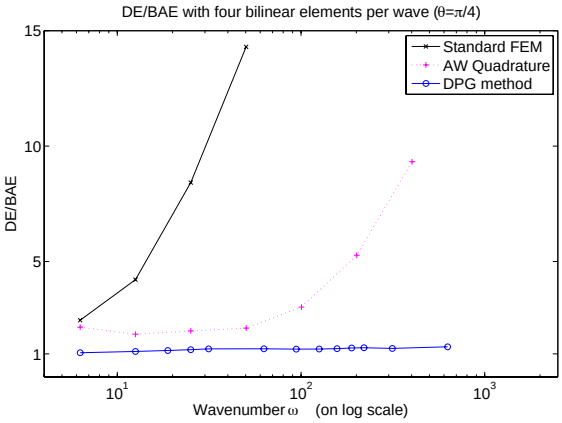
$$\hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{u} = f, \qquad \text{on } \Omega \tag{41b}$$

$$\vec{u}\cdot\vec{n} = u_n, \qquad \text{on } \Gamma_1 \cup \Gamma_4 \tag{41c}$$

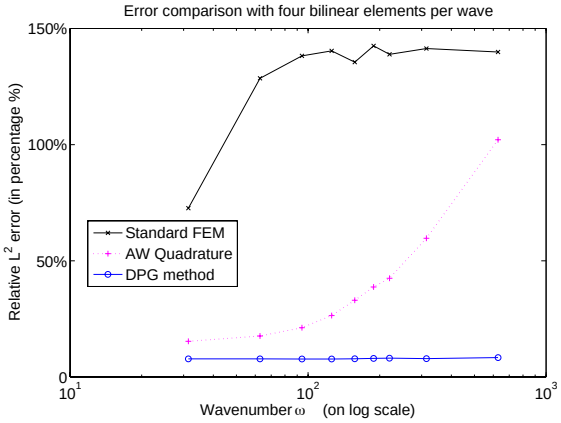
$$\vec{u}\cdot\vec{n} - \phi = g, \qquad \text{on } \Gamma_2 \cup \Gamma_3, \tag{41d}$$

i.e., we prescribe the normal "velocity" on the lower and left edges of the domain ($\Gamma_1 \cup \Gamma_4$) and maintain the previous Robin boundary conditions at the upper and right edges ($\Gamma_2 \cup \Gamma_3$). We choose f, u_n, and g so that the exact solution is the same as in Model Problem A.

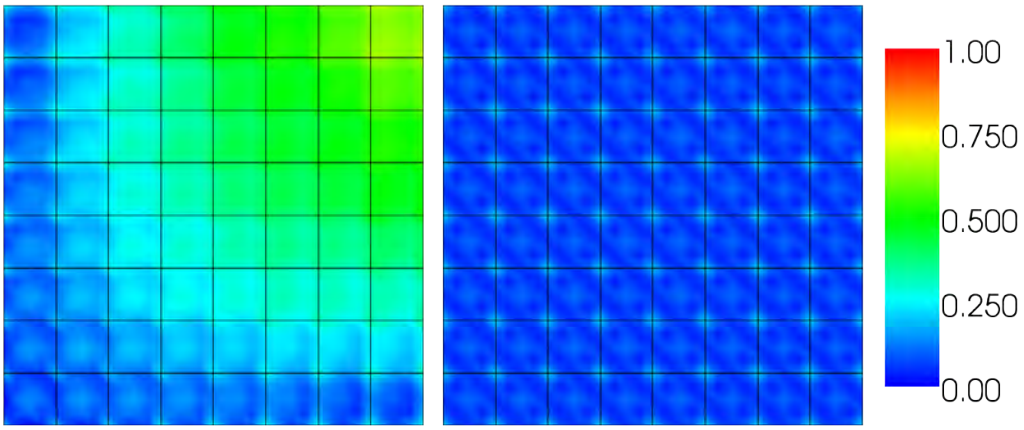
The results are depicted in Figure 2. In Figure 2(a), we observe a behavior similar to Figure 1. In Figure 2(b), we see that the relative L^2 error percentage remains more or less constant (about 8%) for the DPG method, while it increases with the number of degrees of freedom to about 140% for the standard method. (Note that, as before, the number of degrees of freedom is tied to the wavenumber through $\omega h = \pi/2$.)



(a) The DE/BAE ratio for the three methods



(b) Relative L^2 error (in %) for the three methods



(c) *Left:* Plot of the error for the CG method showing accumulation of phase errors (at the northeast corner) in the direction of wave propagation. *Right:* Plot of the error for the DPG method showing no accumulation of phase errors in the propagation direction.

FIGURE 2. Results for Model B (all results are for $\theta = \pi/4$)

5.3. **Model problem C.** In this problem, the exact solution consists of the cylindrical wave

$$\phi(\vec{x}) = H_0^{(2)}(\omega|\vec{x}|),$$

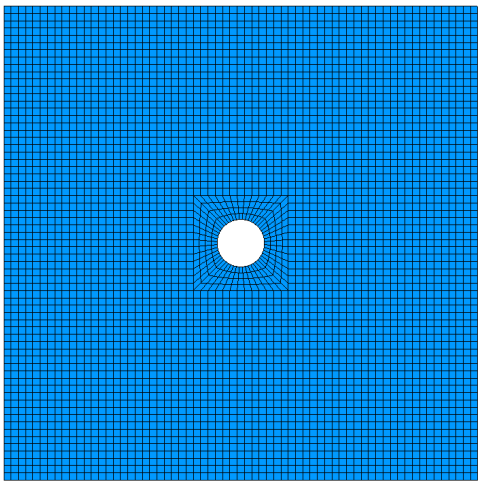
where $H_0^{(2)}$ is the zero-order Hankel function of the second kind. The domain consists of the square $(-1,1)^2$ with a circular exclusion of radius $a = 0.1$ in the center, i.e. $\Omega = (-1,1)^2 \setminus \{\vec{x} : |\vec{x}| \leq a\}$. We denote the boundary of the circle by Γ_a. The equations of the boundary value problem are

$$\hat{\imath}\omega\vec{u} + \vec{\nabla}\phi = \vec{0}, \qquad \text{on } \Omega \tag{42a}$$

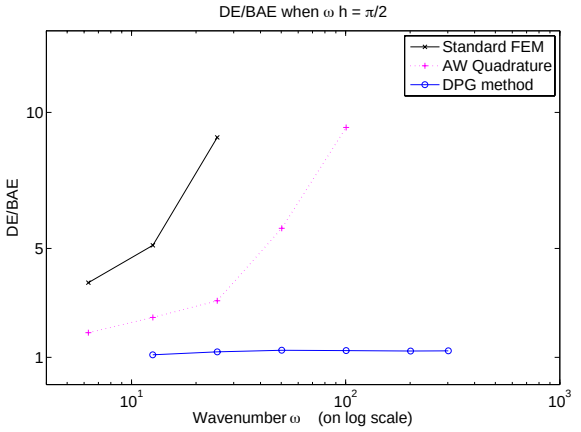
$$\hat{\imath}\omega\phi + \vec{\nabla}\cdot\vec{u} = 0, \qquad \text{on } \Omega \tag{42b}$$

$$\vec{u}\cdot\vec{n} = u_n, \qquad \text{on } \Gamma_a \tag{42c}$$

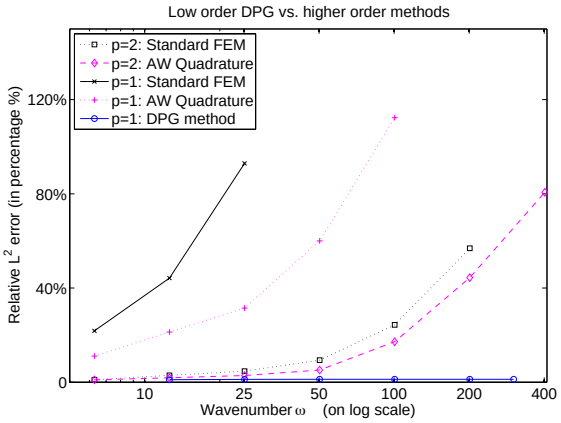
$$\vec{u}\cdot\vec{n} - \phi = g, \qquad \text{on } \partial\Omega \setminus \Gamma_a. \tag{42d}$$



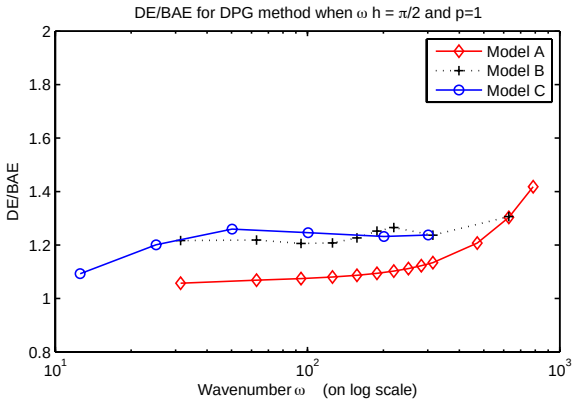
(a) Mesh when $\omega = 16\pi$. Note the use of curved elements.



(b) Ratio of the discretization error to the best approximation error (in L^2-norm) for three methods.



(c) The DPG method, at its lowest order ($p = 1$), compares favorably to other higher order ($p = 2$) methods.



(d) A comparison of the performance of DPG method for this model and the previous two model problems (with zoomed ordinate scale).

FIGURE 3. Results for Model C.

We discretize using a grid of square elements, with the exception of a block of elements which are deformed by geometry mappings generated through transfinite interpolation – see Figure 3(a).

The results are similar to the previous two model problems. The DPG error closely follows the L^2 best approximation error, as evidenced by the ratio plotted in Figure 3(b), which remains close to the optimal value of 1. These results are for $p = 1$.

We also compared these results with the $p = 2$ case of the standard FEM. It is well known that increasing the order improves pollution errors for the standard method. This is indeed the case, as seen in Figure 3(c), for both the standard FEM and its modification of [1]. However, even with this improved performance, neither of these methods compared favorably to the *lowest* order DPG method. While the DPG error remained under 9% (in the L^2 norm) for the range of wave numbers considered, the L^2 errors of the other methods eventually increased beyond a 100%.

Finally, in Figure 3(d), to get an idea of the relative difficulty of the model problems we considered so far, we compared the performance of the DPG method for Models A, B, and C. The DE/BAE ratio remains close to the optimal value of one for all the three cases. Model A, run with the propagation direction $\theta = 0$ in this plot, seems to show the largest increase in DE/BAE as the wavenumber increases. (We have run this model to the limit of our computational resources.) Although this increase is a small fraction of the increase the other methods suffer from, the fact that there is a slight increase seems to indicate that the DPG method may also suffer from pollution errors for large enough wave numbers. The current data however is insufficient to make a definitive conclusion.

5.4. **Model problem D: Pekeris waveguide.** Finally, we consider a more realistic example of wave propagation. The Pekeris waveguide (see Figure 4(a)) is a canonical example of a shallow water waveguide. This model consists of a water layer above a sediment layer. A point source within the water column at depth z_s generates time-harmonic pulses which propagate into the water and sediment layers. The sediment layer is represented as an acoustic medium with higher density and sound speed. The change in acoustic properties occurs at a depth H. At the surface Γ_1 of the waveguide, a pressure-release boundary condition is prescribed (i.e. $\phi = 0$). The speed of sound in water is taken to be $c = 1500$ meters per second. We set $L = 1500$ m to be our length scale in non-dimensionalization of the problem. An additional scaling is applied to the ambient density so that $\rho_0 = 1$ within the water column. The result of this scaling is such that $\vec{u}$ and ϕ are of the same order of magnitude. The full set of problem parameters after non-dimensionalization is as follows:

$c_1 = 1$	speed of sound in water
$c_2 = 1.2$	speed of sound in sediment
$\rho_1 = 1$	ambient density of water
$\rho_2 = 1.8$	ambient density of sediment
$H = 1/15$	depth of water column
$z_s = 36/1500$	depth of point source.

The original Pekeris problem is posed on the unbounded half space $\{(x, z) \subset \mathbb{R}^2 : z < 0\}$ with Sommerfeld radiation conditions. This can be numerically solved using PML or other techniques for truncating infinite domain problems. However, since these truncation techniques are not the subject of the present study, we construct a simpler model problem by truncating to the domain $\Omega = [0, 200/L] \times [0, H/L]$ and simply imposing non-homogeneous Robin boundary conditions with contrived data obtained from the exact solution:

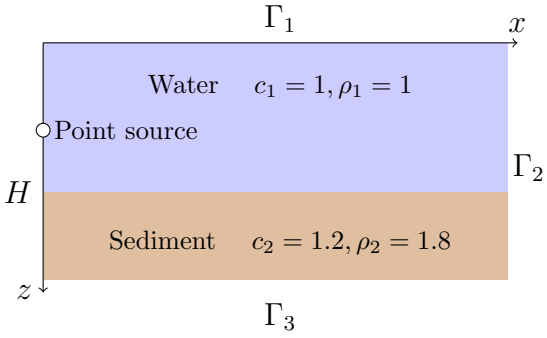
$$\frac{i\omega}{\rho_0 c^2}\phi + \nabla \cdot \vec{u} = \delta_{z_s}, \qquad \text{in } \Omega \tag{43a}$$

$$i\omega\rho_0\vec{u} + \nabla\phi = 0, \qquad \text{in } \Omega \tag{43b}$$

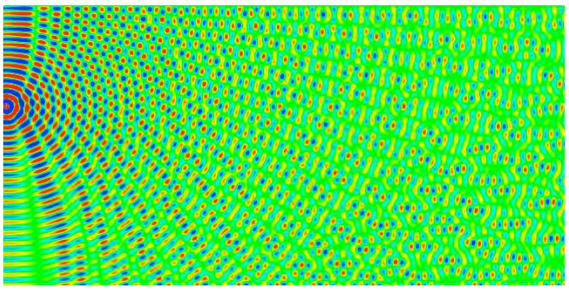
$$\phi = 0, \qquad \text{on } \Gamma_1 \tag{43c}$$

$$\vec{u} \cdot \vec{n} - \phi = g, \qquad \text{on } \Gamma_2 \tag{43d}$$

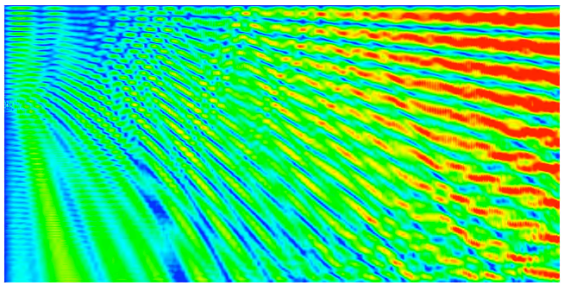
$$\vec{u} \cdot \vec{n} = u_n, \qquad \text{on } \Gamma_3 \tag{43e}$$



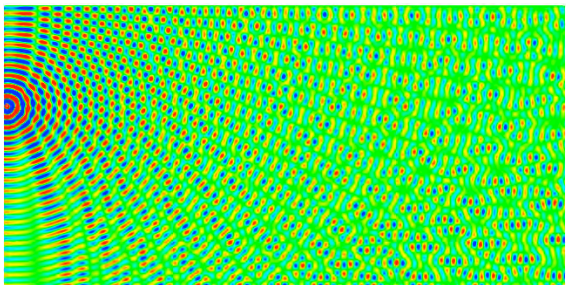
(a) Diagram of the Pekeris waveguide.



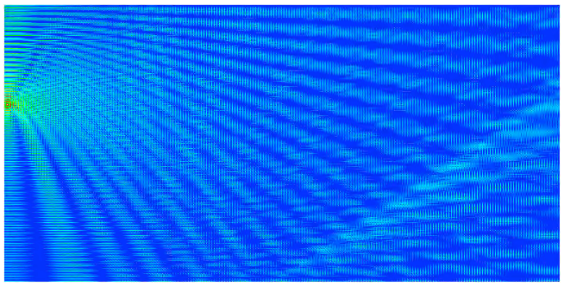
(b) The exact solution for $\omega = 2\pi \times 400$.



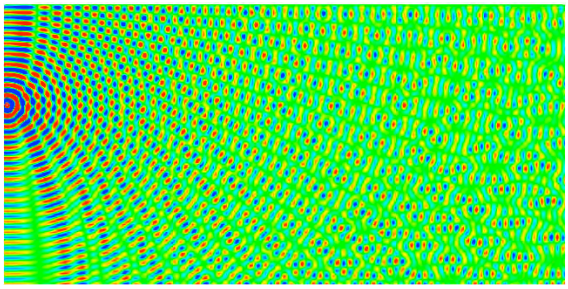
(c) Error from standard biquadratic FEM using about 4 elements per wavelength ($\omega = 2\pi \times 400$).



(d) Solution from standard biquadratic FEM using about 4 elements per wavelength ($\omega = 2\pi \times 400$).



(e) Error from DPG method using about 4 bilinear elements per wavelength ($\omega = 2\pi \times 400$).



(f) Solution from DPG method with 4 bilinear ($p = 1$) elements per wavelength ($\omega = 2\pi \times 400$).

FIGURE 4. Model D: Pekeris waveguide. Only the real parts of the scalar component are shown. The color scale for Figures 4(b), 4(d), and 4(f) is $[-10 \cdots 10]$, while for Figures 4(c) and 4(e), it is $[0 \cdots 5]$.

An analytic expression for the pressure ϕ within the water column, which can be derived through Fourier transformation and complex-contour integration (see, e.g. [19]), is given by

$$\phi(x,z) = -2i \int_{\sqrt{k_1^2 - k_2^2}}^{\infty} \frac{\gamma_1 \gamma_2 \sin(\gamma_1 z_s) \sin(\gamma_1 z)}{\gamma_1^2 \frac{\rho_2}{\rho_1} \cos^2(\gamma_1 H) + \gamma_2^2 \frac{\rho_1}{\rho_2} \sin^2(\gamma_1 H)} e^{-i\beta x} d\gamma_1 + \frac{-i\pi}{\rho(z_s)} \sum_{m=1}^{N} u_m(z_s) u_m(z) e^{-i\beta_m x}, \tag{44}$$

where

$$\begin{cases} k_1 = \omega/c_1, \\ k_2 = \omega/c_2, \\ \beta = \sqrt{k_1^2 - \gamma_1^2} > 0, \\ \gamma_2 = \sqrt{k_2^2 - \beta^2} > 0, \end{cases}$$

and the N modal values $\beta_m = \sqrt{k_1^2 - \gamma_{1m}^2} = \sqrt{k_2^2 - \gamma_{2m}^2} > 0$ are determined by the N solutions of the characteristic equation

$$\tan(\gamma_{1m} H) = \frac{i\gamma_{1m}\rho_2}{\gamma_{2m}\rho_1}.$$

A plot of the exact solution is shown in Figure 4(b).

Plots of the errors and computed solutions are in Figure 4. Since the results from the lowest order (bilinear) case of the FEM are poor (not displayed), we only compare the bilinear DPG solution to the biquadratic FEM. A close observation of the solution computed using the standard FEM in Figure 4(d) shows a small phase lag when compared to the exact solution in Figure 4(b). Since this may be hard to judge from Figure 4(b), we also include visualization of the errors in Figures 4(c) and 4(e). Clearly, away from the source, the error is large for the standard method, while for the DPG method, the error remains more or less the same throughout the domain. Note also that small phase errors can lead to large L^2 errors. The DPG solution, even in the lowest order case, shown in Figure 4(f), is a far better approximation (visually almost identical to the exact solution).

6. Conclusion

6.1. **Summary.** We presented a new DPG method for acoustic time harmonic wave propagation. Although this method has more unknowns than other standard methods, and although it does not have conservation properties in its current form, we think it is an interesting alternative because it exhibits remarkably small phase errors in all attempted numerical experiments. While many standard methods show comparable performance for low to moderate wave numbers, for large wave numbers, the new method is highly competitive. Our analysis using the known regularity and stability results for the Helmholtz equation leads to a proof of error estimates, which however does not explain the low phase errors.

6.2. **The analysis in hindsight.** One may observe, as we did in hindsight, that the analysis we performed in Section 4 has elements that can be generalized to apply for various problems beyond wave propagation. To briefly remark on a way to generalize, consider any abstract problem $Au = f$, with a linear operator $A : D(A) \mapsto L^2(\Omega)$, where the domain of A, $D(A)$, incorporates any boundary conditions on u, and is equipped with the graph norm $\|u\|_{D(A)}^2 = \|u\|_\Omega^2 + \|Au\|_\Omega^2$. Formally introducing an L^2 adjoint operator A^* by $(Au, v)_\Omega = (u, A^*v)_{\Omega_h} + \langle\!\langle u, v \rangle\!\rangle_{\partial\Omega_h}$, where we have lumped all element interface and boundary terms into $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{\partial\Omega_h}$, we can pose an abstract ultraweak formulation as follows: Find u in $L^2(\Omega)$ satisfying $(u, A^*v)_{\Omega_h} + \langle\!\langle \hat{u}, v \rangle\!\rangle_{\partial\Omega_h} = (f, v)_{\Omega_h}$ for all v in the space V with norm defined by $\|v\|_V^2 = \|v\|_\Omega^2 + \|A^*v\|_{\Omega_h}^2$. If we set A to the Helmholtz (first order) wave operator, we find that this is exactly the formulation on which the DPG method of this paper is based, cf. (13). Note that in the above generalization, $\hat{u}$ is sought in a space Q, which is normed by a straightforward generalization of the quotient norms in § 2.4. We

can then view the entire analysis of Section 4 as aimed at proving the equivalence of the $\|\cdot\|_V$-norm with the "optimal norm"

$$\|v\|^2_{\mathrm{opt},V} = \|A^*v\|^2_{\Omega_h} + \big|[v]\big|^2_{\partial\Omega_h}, \quad \text{where} \quad \big|[v]\big|_{\partial\Omega_h} = \sup_{u\in D(A)} \frac{\langle\!\langle u, v\rangle\!\rangle_{\partial\Omega_h}}{\|u\|_{D(A)}},$$

cf. §4.2. The inequality $\|v\|_{\mathrm{opt},V} \le C\|v\|_V$ can be proved, even in the general context, exactly as in proof of the upper bound in Theorem 4.5. The gist of the argument to prove the reverse inequality can be abstracted from Section 4, under the assumption that $\|u\|_\Omega \le C_0\|Au\|_\Omega$. (This assumption follows from Lemmas 4.2 and 4.3 in our particular case.) Then, given any $v \in V$, considering a u that solves $Au = v$, we have

$$\begin{aligned}
\|v\|^2_\Omega &= (Au, v)_\Omega = (u, A^*v)_{\Omega_h} + \langle\!\langle u, v\rangle\!\rangle_{\partial\Omega_h} \\
&\le \|u\|_\Omega \|A^*v\|_{\Omega_h} + \|u\|_{D(A)}\big|[v]\big|_{\partial\Omega_h} \\
&\le \left(C_0^2\|v\|^2_\Omega + (C_0^2+1)\|v\|^2_\Omega\right)^{1/2} \left(\|A^*v\|^2_{\Omega_h} + \big|[v]\big|^2_{\partial\Omega_h}\right)^{1/2},
\end{aligned}$$

which proves that $\|v\|_V \le C\,\|v\|_{\mathrm{opt},V}$, thus completing the proof of the norm equivalence. In hindsight, we understand that this is the essence of the analysis not only in this paper, but also in [6].

6.3. **Future directions.** Our future studies are focused on sharpening our theoretical tools and formalizing and improving the informally stated abstract generalization above (§ 6.2), all with the aim of proving or disproving that the errors are pollution-free. Another direction of research we are exploring exploits the Hermitian positive definiteness of the linear system, together with the good phase errors, to design efficient iterative solvers.

Acknowledgements. We are grateful to Markus Melenk for valuable discussions on the theory in this paper and to David Pardo and Victor M. Calo for constructive criticisms on our numerical experiments.

References

[1] M. Ainsworth and H. A. Wajid, *Optimally blended spectral-finite element scheme for wave propagation and nonstandard reduced integration*, SIAM Journal on Numerical Analysis, 48 (2010), pp. 346–371.

[2] I. M. Babuška and S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM Rev., 42 (2000), pp. 451–484 (electronic). Reprint of SIAM J. Numer. Anal., 34 (1997), no. 6, pp. 2392–2423.

[3] A. Brandt and I. Livshits, *Wave-ray multigrid method for standing wave equations*, Electron. Trans. Numer. Anal., 6 (1997), pp. 162–181 (electronic). Special issue on multilevel methods (Copper Mountain, CO, 1997).

[4] O. Cessenat and B. Despres, *Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem*, SIAM J. Numer. Anal., 35 (1998), pp. 255–299 (electronic).

[5] R. Courant and K. O. Friedrichs, *Supersonic Flow and Shock Waves*, Interscience Publishers, Inc., New York, N. Y., 1948.

[6] L. Demkowicz and J. Gopalakrishnan, Analysis of the DPG method for the Poisson equation. *SIAM J Numer. Anal.*, 49(5):1788–1809, 2011.

[7] L. Demkowicz and J. Gopalakrishnan, *A class of discontinuous Petrov-Galerkin methods. Part I: The transport equation*, Computer Methods in Applied Mechanics and Engineering, 199 (2010), pp. 1558–1572.

[8] ——, *A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions*, Numerical Methods for Partial Differential Equations, 27 (2011), pp. 70–105.

[9] L. Demkowicz, J. Gopalakrishnan, and A. Niemi, *A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity*, To appear in Applied Numerical Mathematics (2011).

[10] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl *Polynomial extension operators. Part III*, To appear in Math. Comp., (2011).

[11] L. F. Demkowicz, *Polynomial exact sequences and projection-based interpolation with application to maxwell equations*, in Mixed finite elements, compatibility conditions, and applications, vol. 1939 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 2008, pp. x+235. Lectures given at the C.I.M.E. Summer School held in Cetraro, June 26–July 1, 2006, Edited by Boffi and Lucia Gastaldi.

[12] X. Feng and H. Wu, *Discontinuous Galerkin methods for the Helmholtz equation with large wave number*, SIAM J. Numer. Anal., 47 (2009), pp. 2872–2896.

[13] X. Feng and H. Wu, *hp-discontinuous Galerkin methods for the Helmholtz equation with large wave number*, To appear in Math. Comp., (2010).

[14] J. Gopalakrishnan and M. Oh, *Commuting smoothed projectors in weighted norms with an application to axisymmetric Maxwell equations.* J. Scientific Computation, DOI: 10.1007/s10915-011-9513-3, 2011.

[15] J. Gopalakrishnan and W. Qiu, An analysis of the practical DPG method. *Submitted*, 2011.

[16] R. Hiptmair, A. Moiola, and I. Perugia, *Plane wave discontinuous Galerkin methods for the 2D Helmholtz equation: analysis of the p-version*, Tech. Rep. 2009-20, Eidgenössische Technische Hochschule, 2009.

[17] T. Huttunen, P. Monk, and J. P. Kaipio, *Computational aspects of the ultra-weak variational formulation*, J. Comput. Phys., 182 (2002), pp. 27–46.

[18] F. Ihlenburg, *Finite element analysis of acoustic scattering*, vol. 132 of Applied Mathematical Sciences, Springer-Verlag, New York, 1998.

[19] F. B. Jensen, W. A. Kuperman, M. B. Porter, and H. Schmidt, *Computational Ocean Acoustics*. AIP Press, 2000.

[20] B. Lee, T. A. Manteuffel, S. F. McCormick, and J. Ruge, *First-order system least-squares for the Helmholtz equation*, SIAM J. Sci. Comput., 21 (2000), pp. 1927–1949. Iterative methods for solving systems of algebraic equations (Copper Mountain, CO, 1998).

[21] J. M. Melenk, *On Generalized Finite Element Methods*, PhD thesis, University of Maryland, 1995.

[22] J. M. Melenk and S. Sauter, Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Math. Comp*, 79:1871–1914, 2010.

[23] R. Tezaur and C. Farhat, *Three-dimensional discontinuous Galerkin elements with plane waves and Lagrange multipliers for the solution of mid-frequency Helmholtz problems*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 796–815.

[24] K. Yosida, *Functional analysis*, Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

[25] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. Calo, *A class of discontinuous Petrov-Galerkin methods. Part IV: Wave propagation*, Journal of Computational Physics, 230 (2011), pp. 2406–2432.

Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712.
E-mail address: leszek@ices.utexas.edu

724 SW Harrison St., Portland State University, Portland, OR 97201.
E-mail address: gjay@pdx.edu

Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Casilla 4059, Valparaiíso, Chile.

Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712.
E-mail address: jzitelli@ices.utexas.edu